# An Automated Stopping Rule for MCMC Convergence Assessment

Didier Chauveau, Jean Diebolt

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

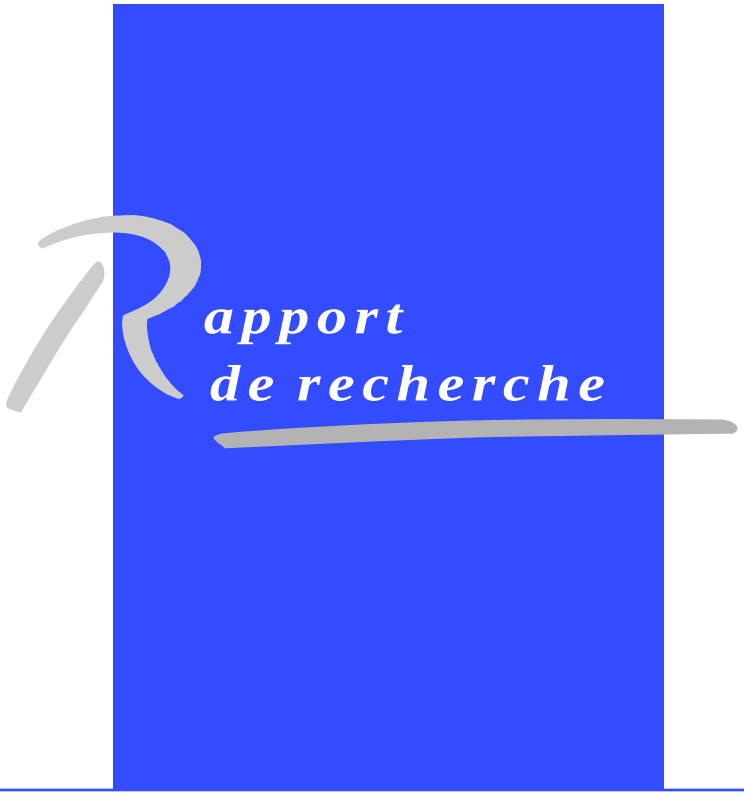
# *An automated stopping rule for MCMC convergence assessment*

Didier Chauveau and Jean Diebolt

**No 3566**

Novembre 1998

THÈME 4




ISSN 0249-6399

# An automated stopping rule for MCMC convergence assessment

Didier Chauveau and Jean Diebolt

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet is2

Rapport de recherche n° 3566 — Novembre 1998 — 33 pages

**Abstract:** In this paper, we propose a methodology essentially based on the Central Limit Theorem for Markov chains to monitor convergence of MCMC algorithms using actual outputs. Our methods are grounded on the fact that normality is a testable implication of sufficient mixing. The first control tool tests the normality hypothesis for normalized averages of functions of the Markov chain over independent parallel chains started from a dispersed distribution. A second connected tool is based on graphical monitoring of the stabilization of the variance after $n$ iterations near the limiting variance appearing in the CLT. Both methods work without knowledge on the sampler driving the chain, and the normality diagnostic leads to automated stopping rules. The methodology is developed for finite state Markov chains, and extended to the continuous case. Heuristic procedures based on Berry-Esséen bounds are also investigated. These stopping rules are implemented in a software toolbox whose performances are illustrated through simulations for finite and continuous state chains reflecting some typical situations (slow mixing, multimodality) and a full scale application. Comparisons are made with the binary control method of Raftery and Lewis.

**Key-words:** asymptotic variance, convergence assessment, finite state Markov chain, MCMC algorithm, normality test, stationarity, CLT for Markov chains

*(Résumé : tsvp)*

Didier Chauveau, Université de Marne-la-Vallée, Analyse et Mathématiques Appliquées, 5 Bd Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2, France. <chauveau@math.univ-mlv.fr>
Jean Diebolt, CNRS, UMR 5523–LMC, Équipe de Statistique et de Modélisation Stochastique, BP 53, 38041 Grenoble Cedex 09, France. <Jean.Diebolt@imag.fr>

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France)
Téléphone : 04 76 61 52 00 - International: +33 4 76 61 52 00
Télécopie : 04 76 61 52 52 - International: +33 4 76 61 52 52

# Une règle d'arrêt automatique pour le contrôle de convergence des algorithmes MCMC

**Résumé :** Nous proposons une méthodologie de contrôle des algorithmes MCMC fondée sur le Théorème de Limite Centrale (TLC) pour les chaînes de Markov. La normalité est en effet une conséquence vérifiable du fait que la chaîne a suffisamment visité le support de la loi cible. Le premier outil proposé teste la normalité d'un échantillon de sommes de fonctions de la chaîne construit à partir de chaines parallèles initialisées suivant une loi suffisamment dispersée. Une seconde technique naturellement liée à la première consiste à contrôler la stabilisation de la variance asymptotique intervenant dans le TLC. Cette approche conduit à une méthode de contrôle non spécifique de l'algorithme MCMC étudié, pour laquelle nous proposons des critères d'arrêt automatiques. Le cas des chaînes à espace fini est d'abord étudié, car il permet un développement précis du contrôle de la stabilisation de la variance et la comparaison avec des techniques utilisant l'inégalité de Berry-Esséen. L'extension aux chaînes générales est ensuite proposée. L'aspect générique de la méthode a justifié la réalisation d'un logiciel de contrôle qui implémente ces critères de convergence. Ce logiciel est utilisé ici pour illustrer la pertinence de ces outils de contrôle dans diverses situations typiques, discrètes ou continues (chaînes à faible mélangeance, lois multimodales, applications réelles). Notre méthode est à chaque fois comparée avec le contrôle binaire de Raftery et Lewis.

**Mots-clé :** Algorithmes MCMC, chaînes de Markov discrètes, contrôle de convergence, TLC pour les chaînes de Markov, stationnarité, variance asympotique

# 1 Introduction

Markov Chain Monte Carlo (MCMC) algorithms, introduced by Gelfand and Smith (1990), generate ergodic Markov chains $(x^{(t)})$ with state space $E$ and invariant probability measure $\pi$ for which direct simulation is not tractable (i.e. independent random variables distributed according to $\pi$ cannot be simulated), or for which computation of integrals of the form

$$\int_E f(x)\,\pi(dx) \tag{1}$$

cannot be achieved. Since $\pi$ is the only invariant probability measure of the ergodic Markov chain $(x^{(t)})$, the $x^{(T+t)}$'s, $t \geq 0$ are approximately $\pi$ distributed for $T$ large enough and (1) is approximated by

$$\frac{1}{n}\sum_{t=1}^{n} f(x^{(T+t)}). \tag{2}$$

Objectives of the end user are to gather information about the precision in the approximation of (1) by (2), to provide a detailed picture of $\pi$ and in some situations to output an iid sample from $\pi$.

Convergence control (or diagnosis) techniques have been addressed to answer such questions, and several methods have been proposed in the recent literature (see, e.g., Brooks and Roberts 1995, and Robert 1996, for a survey). These diagnostics can be based upon one single output (*single chain*) or upon outputs from several independent replications of the chain started from a preassigned initial distribution (*parallel chains*). An important criterion is the computer investment: diagnostics requiring problem-specific computer codes for their implementation (e.g., requiring knowledge of the transition kernel of the Markov chain) are far less usable for the end user than diagnostics solely based upon the outputs from the sampler. The latter can use available generic code. Last but not least, interpretability is important. As Brooks and Roberts (1995) point out, "a diagnostic which produces a definitive solution will generally be preferred to one which requires subjective interpretation and/or experience on the part of the user".

Both parallel and single chain methods have well-known advantages and drawbacks (Brooks and Roberts 1995, or Robert 1996), but we believe that only parallel chain methods can provide satisfactory control tools. Although parallel methods obviously require a larger computational expense, they are more dedicated to output iid random variables from $\pi$ and convey more confidence that the whole support of $\pi$ has been explored (see Gelman and Rubin 1992 for a discussion). Above all, checking convergence to stationarity of $(x^{(t)})$ basically requires comparing the distributions of $x^{(t)}$ for different values of $t$. This involves comparing probabilities of $x^{(t)}$-measurable events for different values of $t$. Such probabilities are limits of occurrence frequencies of these events for sequences of independent identical experiments. Therefore, they can reasonably be evaluated only through several independent sequences started from a same initial distribution.

In this paper, we propose a new methodological approach for assessing convergence of MCMC algorithms. Our approach is grounded on the fact that normality is an implication

of sufficient mixing, which is testable across parallel sequences issued from a dispersed initial distribution, and allows for controlling precision. Hence, instead of checking for stationarity of $(x^{(t)})$, we primarily aim at controlling the precision of estimates like (2). A natural way to do this is through confidence regions based on normal approximation resulting from the Central Limit Theorem (CLT) for Markov chains. Difficulties arise since we have to make use of two asymptotic results (as $n \to \infty$), the CLT and the convergence to the limiting variance. This is the reason why we propose, first, to use statistical tests for testing normality of the normalized sums

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left( h(x^{(t)}) - \pi h \right), \quad \pi h = \int_E h(x)\, \pi(dx), \tag{3}$$

using samples obtained from parallel chains, and second, to monitor variance stabilization. Our approach results in control techniques which comply with the above criterion, i.e. they are not problem-specific (hence a *generic* computer code has been developed and is publicly available), and they provide automated diagnostics. The ideas are first presented thoroughly in the case of finite state Markov chains. Our motivations for adopting this point of view are first that more precise mathematical tools exist for finite chains. In particular, the limiting variance in the CLT can be consistently estimated and compared to an estimate of the variance after $n$ iterations. Also, the stationary probabilities $\pi_i$ for each state $i$ can be estimated together with confidence intervals. Finally, the proposed control methods can be applied to large finite chains resulting from actual situations or to finite chains obtained from continuous state Markov chains through a theoretically valid *discretization* procedure (Guihenneuc and Robert 1998), or through the *duality principle* (Diebolt and Robert 1994).

A single chain technique making use of finite Markov chain theory has already been proposed by Raftery and Lewis (1992, 1996). Their *binary control* relies on an approximation of some binary process issued from a general MCMC algorithm by a two-state Markov chain. However, this approximation is rather weak (see, e.g., Robert 1996). This method is nevertheless one of the most popular and commonly used, mainly because it is not problem-specific, delivers an automated stopping rule and is available in existing software libraries (STATLIB). These are the reasons why we will compare our approach against this competing method.

Section 2 contains the theoretical background for finite ergodic Markov chains which will be used in the paper. Connections between the CLT and the renewal theory are recalled. The main tool of this section is a CLT for the time spent in a state during the first $n$ steps of an ergodic Markov chain, with the limiting variance available through algebraic computations involving the transition matrix and the invariant probability (Kemeny and Snell 1960). A heuristic procedure for MCMC convergence assessment using Berry-Esséen type error bounds in the CLT (Feller 1968) is also investigated. Section 3 describes the two control methods we propose for finite Markov chains: A test of normality for (3) with indicator functions $f = \mathbb{I}_i$, $i \in E$, and a comparison, using graphical monitoring, between consistent estimates of the limiting variance and of the variance after $n$ iterations. Section 4 extends this methodology to continuous state space Markov chains. The extension of the normality monitoring to the general case is almost straightforward, and automated stopping

rules are proposed. The variance comparison is also carried out, at the expend of some approximation for estimation of the limiting variance. Section 5 is devoted to illustrative examples and comparisons with the binary control, which is briefly described there. A toy example for a finite state Markov chains is then studied. It shows that the normality control is a powerful tool for detecting slowly mixing chains, and that our parallel chains method is preferable to — and faster than — a single chain method. Comparisons between the time needed to reach approximate stationarity and the time needed to accept normality hypotheses are given on a second example based on a random walk over the $d$-dimensional cube. Examples for continuous state MCMC algorithms, arising from typically illustrative situations (multimodal posterior), and actual applications (mixture of distributions) are also proposed.

# 2 CLT and renewal theory in the discrete case

The beginning of this section contains classical theoretical result for finite state Markov chains which will be used in the paper. We consider a finite irreducible aperiodic Markov chain $x^{(t)}$ with finite state space $E$, $|E| = K$, transition matrix $\mathbb{P}$ and invariant probability $\pi = (\pi_i, i \in E)$.

## 2.1 Renewal times

For each subset $A$ of $E$, we denote by $N_n(A) = \sum_{t=1}^{n} \mathbb{I}\left(x^{(t)} \in A\right)$ the occupation time of $A$ during the first $n$ steps. For each real function $h$ defined on the state space, consider

$$S_n(h) = \sum_{t=1}^{n} h\big(x^{(t)}\big) \quad \text{and} \quad S_n(\bar{h}) = \sum_{t=1}^{n} \Big[h\big(x^{(t)}\big) - \mathbb{E}^{\pi}\left[h\right]\Big].$$

We assume for simplicity that the Markov chain starts from $x^{(0)} = i$. When $x^{(0)}$ is generated from an initial distribution $\mu_0$, we only have to shift the starting time to the first time $x^{(t)} = i$. Let $T_i(1) = \inf\{t > 0 : x^{(t)} = i\}$ be the first time $t > 0$ the chain returns to the state $i$, and $T_i(0) = 0$ by convention. The r.v. $T_i(1)$ is a stopping time with respect to the sequence $(x^{(t)})_{t \geq 0}$. Define the stopping time $T_i(p)$, $p \geq 2$, as the $p$th return time to state $i$. Let $\tau_i(p)$, $p \geq 1$, be the duration of the $p$th excursion out of state $i$. The $\tau_i(p)$'s and $T_i(p)$'s are connected by $T_i(1) = \tau_i(1)$ and $T_i(p) = T_i(p-1) + \tau_i(p), \quad p \geq 1$.

**Proposition 1** *For any $i \in E$, the $\tau_i(p)$'s, $p \geq 1$, are iid and have finite moments of all orders. Moreover, $\mathbb{E}_i[\tau_i(1)] = \mathbb{E}_{\mu_0}[\tau_i(p)] = \pi_i^{-1}$ for $p \geq 2$.*

Proposition 1 can be found in Chung (1967). Note that it holds for any starting distribution by considering only the $\tau_i(p)$'s for $p \geq 2$. Let $q_i(t)$ be the random number of returns to state $i$ before time $t$, $q_i(t) = \max\{p \geq 1 : T_i(p) \leq t\}$. We have $q_i(t) + 1 = \sum_{s=0}^{t} \mathbb{I}\big(x^{(s)} = i\big)$, from

which it follows that $\mathbb{E}_i\left[q_i(t)+1\right] = \sum_{s=0}^{t} p_{ii}^{(s)}$, where $\mathbb{P}^s = \left(p_{ii}^{(s)}\right)$. Therefore,

$$\lim_{t\to\infty} \mathbb{E}_i\left[\frac{q_i(t)+1}{t+1}\right] = \pi_i,$$

and a consequence of the strong law of large numbers for ergodic Markov chains is that

$$\lim_{n\to\infty} \frac{N_n(i)}{n} = \lim_{n\to\infty} \frac{q_i(n)+1}{n+1} = \pi_i \quad \text{a.s.} \tag{4}$$

Finally, we define for $p \geq 0$ the block sums over the excursions out of $i$:

$$Z_p(h) = \sum_{t=T_i(p)+1}^{T_i(p+1)} h\left(x^{(t)}\right) \quad \text{and} \quad Z_p(\bar{h}) = \sum_{t=T_i(p)+1}^{T_i(p+1)} \left[h\left(x^{(t)}\right) - \mathbb{E}^{\pi}\left[h\right]\right].$$

**Proposition 2** *Let the finite state Markov chain $(x^{(t)})_{t\geq 0}$ start from $x^{(0)} = i$. Then for any $h$ the $Z_p(h)$'s, $p \geq 0$, are iid random variables and have finite moments of all orders. Moreover, $\mathbb{E}_{\mu_0}\left[Z_p(h)\right] = \pi h/\pi_i$ for $p \geq 1$.*

Proposition 2 can be found in Chung (1967). It also holds for any starting distribution by considering the $Z_p(h)$'s for $p \geq 1$.

## 2.2 CLT and limiting variance

We will make use of Wald's equation (see e.g., Billingsley 1986, p. 306):

**Theorem 1** *Let $Z_1, Z_2, \ldots$ be iid random variables such that $\mathbb{E}[Z_1^2] < \infty$, and $T$ be a stopping time for $(Z_t)_{t\geq 1}$ such that $\mathbb{E}[T] < \infty$. Then*

*(i)* $\mathbb{E}\left[\sum_{p=1}^{T} Z_p\right] = \mathbb{E}[T]\,\mathbb{E}[Z_1]$

*(ii)* $var\left[\sum_{p=1}^{T} Z_p\right] = \mathbb{E}[T]\, var[Z_1]$.

Wald's theorem for square integrable martingales can also be found in Dacunha-Castelle and Duflo (1986, p. 96).

As detailed in Kemeny and Snell (1960), the variance of the random variables $n^{-1/2}$ $S_n(\bar{h})$ converges to a limiting variance

$$\sigma^2(h) = \lim_{n\to\infty} n^{-1} \operatorname{var}_{\mu_0}\left[S_n(\bar{h})\right], \tag{5}$$

which is related to the variance of the $Z_p(h)$'s through the following result:

**Theorem 2** *If the finite Markov chain is irreducible and aperiodic, then for any initial distribution $\mu_0$,*

*(i)* $var_i\,[Z_0(h)] = var_{\mu_0}\,[Z_p(h)] = \dfrac{\sigma^2(h)}{\pi_i} \quad for \quad p \geq 1.$

*(ii)* $\dfrac{S_n(\bar{h})}{\sigma(h)\sqrt{n}} \quad \stackrel{d}{\rightarrow} \quad \mathcal{N}(0,1)$ *as* $n \rightarrow \infty$.

*Proof.* It suffices to prove the result for nonnegative $h$'s and to assume that $x^{(0)} = i$. Since $T_i(q_i(t)) \leq t < T_i(q_i(t)+1)$, it follows that

$$0 \leq T_i(q_i(t)+1) - t < \tau_i(q_i(t)+1).$$

Since $h \geq 0$,

$$\sum_{p=0}^{q_i(t)} Z_p(h) \leq S_t(h) \leq \sum_{p=0}^{q_i(t)+1} Z_p(h).$$

Therefore,

$$\left| \frac{S_t(\bar{h}) - \sum_{p=0}^{q_i(t)} Z_p(\bar{h})}{t} \right| \leq \mathrm{cst} \frac{\tau_i(q_i(t)+1)}{t}, \tag{6}$$

where cst is an appropriate constant. It follows from (5) and (6) that

$$\lim_{t\rightarrow\infty} t^{-1} \mathrm{var}_i \left[ \sum_{p=0}^{q_i(t)} Z_p(\bar{h}) \right] = \lim_{t\rightarrow\infty} t^{-1} \mathrm{var}_i \left[ S_{T_i(q_i(t)+1)}(\bar{h}) \right] = \sigma^2(h). \tag{7}$$

Let $\sigma_Z^2$ denote the common variance of the $Z_p(\bar{h})$'s. The event $\{q_i(t)+1 = n\}$ is measurable for $\big(\tau_i(1), \ldots, \tau_i(n)\big)$ or, equivalently, $\big(Z_0(h), \ldots, Z_{n-1}(h)\big)$-measurable. We apply Wald's equation (Theorem 1) for the iid $Z_p(h)$'s:

$$\begin{aligned} \mathrm{var}_i \left[ Z_0(\bar{h}) + \cdots + Z_{q_i(t)}(\bar{h}) \right] &= \mathrm{var}_i \left[ S_{T_i(q_i(t)+1)}(\bar{h}) \right] \\ &= \mathrm{var}_i \left[ Z_0(\bar{h}) \right] \mathbb{E}_i \left[ q_i(t)+1 \right]. \end{aligned} \tag{8}$$

In view of (4), (7) and (8), we have

$$\sigma^2(h) = \lim_{t\rightarrow\infty} \mathrm{var}_i \left[ \frac{S_{T_i(q_i(t)+1)}(\bar{h})}{t} \right] = \mathrm{var}_i[Z_0(\bar{h})]\,\pi_i,$$

implying *(i)*. The proof of *(ii)* relies on a CLT for a random number of summands (Billingsley 1986, p. 380), applied to the $Z_p(\bar{h})$'s for $1 \leq p \leq q_i(n) - 1$. It makes use of (4) and (6). □

The main tool that we will use in Section 3 is a CLT for the time spent in a given state during the first $n$ steps of an ergodic Markov chain, with the limiting variance available in closed form using $\mathbb{P}$ and $\pi$, as given in Kemeny and Snell (1960). We define two matrices of interest: the matrix $A$ with all rows equal to $\pi$, and the *fundamental matrix*

$$\mathbb{Z} = \big(I - (\mathbb{P} - A)\big)^{-1} = I + \sum_{k=1}^{\infty} (\mathbb{P}^k - A). \tag{9}$$

The limiting variance in the CLT depends on $\mathbb{Z}$ in the following sense: let $h$ and $g$ be two real-valued functions defined on $E$ (considered as column vectors). The limiting covariance matrix is the $K \times K$ symmetric matrix $C = (c_{ij})$ such that, for any starting distribution $\mu_0$,

$$\lim_{n\to\infty} \frac{1}{n} \mathrm{cov}_{\mu_0} \left[ \sum_{t=1}^{n} h(x^{(t)}), \sum_{t=1}^{n} g(x^{(t)}) \right] = h^T C g = \sum_{i,j=1}^{K} h(i) c_{ij} g(j). \tag{10}$$

Note that (10) is stated in Kemeny and Snell (1960) with $\pi$ as the starting distribution to keep computations simple. However, (10) holds for any starting distribution $\mu_0$. The matrix $C$ is related to $\mathbb{Z} = (z_{ij})$ and $\pi$ through

$$c_{ij} = \pi_i z_{ij} + \pi_j z_{ji} - \pi_i \delta_{ij} - \pi_i \pi_j, \tag{11}$$

where $\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ii} = 1$. For each state $i \in E$, let $N_n(i)$ denote, as in §2.1, the occupation time of $i$ during the first $n$ steps. Specializing (10) to the indicator function $h = g = \mathbb{I}_i$ gives the limiting variance

$$\sigma^2(\mathbb{I}_i) = \lim_{n\to\infty} \frac{1}{n} \mathrm{var}_{\mu_0} [N_n(i)] = \mathbb{I}_i^T C \mathbb{I}_i = c_{ii}. \tag{12}$$

The Central Limit Theorem for Markov chains (Theorem 2), when applied to $h = (\mathbb{I}_i, i \in E)$, leads to a multidimensional CLT for the occupation times:

$$\left( \frac{N_n(1) - n\pi_1}{\sqrt{n}}, \ldots, \frac{N_n(K) - n\pi_K}{\sqrt{n}} \right) \xrightarrow{d} \mathcal{N}(0, C).$$

## 2.3 Berry-Esséen bounds for finite Markov chains

In this discrete setting, one purpose of the convergence assessment is to obtain approximate confidence intervals for the $\pi_i$'s. For this, we need to know how large $n$ should be for the normal approximation to be valid. This addresses the question of the convergence rate in the CLT, which naturally leads to the Berry-Esséen theory. In good settings, upper bounds for this rate are given by the Berry-Esséen Theorem for Markov chains, which holds when

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}_{\mu_0} \left[ \frac{S_n(\bar{h})}{\sigma(h)\sqrt{n}} \leq x \right] - \Phi(x) \right| = \mathcal{O}(n^{-1/2}), \tag{13}$$

where $\Phi$ is the standard normal cdf. General conditions have been given for (13) to hold in both the discrete and continuous cases (see Bolthausen 1982). However, a workable bound requires precise estimation of the constant involved in the right-hand side of (13). This question has been investigated by Mann (1996) and Lezaud (1998) for countable state chains, but the proposed constants are far too large for practical use in our case. Moreover, computing these bounds requires knowledge of unavailable quantities (e.g., the *gap* of the transition kernel).

Another approach consists in using the Berry-Esséen Theorem for the iid case (Feller 1971), since in this setup the constant has been precisely evaluated. If $X_i, \ldots, X_n$ are iid random variables with zero expectation, variance $\sigma^2$, and such that $\rho = \mathbb{E}\big[|X|^3\big] < \infty$, then

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left[ \frac{\sum_{t=1}^n X_i}{\sigma\sqrt{n}} \leq x \right] - \Phi(x) \right| < C_{BE} \frac{\rho}{\sigma^3\sqrt{n}},$$

where the constant, initially evaluated at $33/4$, has been lowered down to $C_{BE} \leq 0.7915$ (see, e.g., Seoh and Hallin 1997). This approach can be transposed to the case of Markov chains with the help of renewal theory, through the iid random variables $Z_p(h)$'s defined in §2.1. As a consequence of Proposition 2 and Theorem 2, the distribution of the normalized sum $S_{T_i(q)}(\bar{h})/\sqrt{q}$ converges to a normal distribution with zero mean and variance $\sigma^2(h)/\pi_i$, and the error in the approximation by the normal distribution is given by the Berry-Esséen inequality (for the iid case), provided that $\mathbb{E}\left[|Z_p(h)|^3\right] < \infty$:

$$\begin{aligned} \sup_{x\in\mathbb{R}} \left| \mathbb{P}_{\mu_0}\left[ \frac{\sqrt{\pi_i} S_{T_i(q)}(\bar{h})}{\sigma(h)\sqrt{q}} \leq x \right] - \Phi(x) \right| &\leq \frac{C_{BE}\, \mathbb{E}_{\mu_0}\left[|Z_p(\bar{h})|^3\right]}{\mathrm{var}_{\mu_0}[Z_p(\bar{h})]^{3/2}\sqrt{q}} \\ &\leq \frac{C_{BE} \|\bar{h}\|_\infty^3 \pi_i^{3/2} \mathbb{E}\big[\tau_i^3\big]}{\sigma^3(h)\sqrt{q}}. \end{aligned} \tag{14}$$

The sum $T_i(p) = \sum_{j=1}^p \tau_i(j)$ also satisfies a CLT for iid random variables together with a Berry-Esséen inequality. Defining $\bar{T}_i(q) = \sum_{p=1}^q [\tau_i(p) - 1/\pi_i]$, $\bar{T}_i(q)/\sqrt{q} \xrightarrow{d} \mathcal{N}(0, \mathrm{var}[\tau_i])$ with the error bound

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}_{\mu_0}\left[ \frac{\bar{T}_i(q)}{\sqrt{q\,\mathrm{var}[\tau_i]}} \leq x \right] - \Phi(x) \right| \leq \frac{C_{BE}\, \mathbb{E}\big[|\tau_i - 1/\pi_i|^3\big]}{\mathrm{var}[\tau_i]^{3/2}\sqrt{q}}. \tag{15}$$

**Using the Berry-Esséen bounds for convergence assessment**

A method for MCMC control can be sketched, based on normality assessment for the indicator functions $\mathbb{I}_i$ for $i \in E$, and using Berry-Esséen inequalities to take care of the error. Let us assume that estimates for quantities appearing in upper bounds (14) and (15) are available for some $i \in E$ (essentially we need to estimate $\mathbb{E}[\tau_i^3]$ and $\pi_i$). Then, for a given $\varepsilon_1 > 0$ which represents the acceptable error (the right-hand side of (14)), we can compute the number of "blocks", $q_{i,1} = q_i(\varepsilon_1)$, achieving this error. This choice of $q_{i,1}$ is related to $S_{n_i}(\bar{h})$, where $n_i = \sum_{p=1}^{q_{i,1}} \tau_i(p) \approx q_{i,1}/\pi_i$, and the precision in this approximation by the expectation depends essentially on $\mathrm{var}[\tau_i]$ and the asymptotically normal behavior of $\bar{T}_i(q)$. We can control the error in this last normal approximation using (15) in the same way. For a given error $\varepsilon_2$, this leads to $q_{i,2} = q_i(\varepsilon_2)$. Finally, we need to run the chain up to the observation of $q_i$ returns in $i$, where $q_i = \max(q_{i,1}, q_{i,2})$, to assess normality for both sums. Hence, this control method requires, for state $i$, the simulation of $n_i$ iterations of the chain, where $n_i \approx q_i/\pi_i$ can be estimated together with an approximate confidence interval.

This procedure seems appealing from a theoretical point of view, but unfortunately has two major drawbacks. First, it suffers from the same criticism as many other control methods (e.g., Gelman and Rubin's (1992) variance criterion), since it relies on preliminary estimates of unknown quantities depending on the MCMC algorithm under control itself. Second, it suffers from the poor quality of the standard Berry-Esséen bound. It is known that, even in simple iid situations, the Berry-Esséen bound leads to fairly large sample sizes to ensure that the Kolmogorov-Smirnov distance between the distribution of the normalized sums and the standard normal is smaller than a given $\varepsilon > 0$ (Seoh and Hallin 1997). The simulations in Section 5.2 show that this is also the case in our situation, and that this heuristic leads to dramatically conservative values for the times required to achieve convergence. Hence this procedure is of little practical value, but has been presented here for completeness. We do not discuss it any further (e.g., how to select the states chosen for monitoring; what is the impact of the needed estimates over convergence time) and rather propose empirical methods of control based on normality assessment in the next section.

# 3 Convergence diagnostics with parallel chains

Since we want to use a normal approximation, we need to estimate the time needed to reach approximate normality for suitable functions of $(x^{(t)})$. Intuitively, normality occurs when the parallel chains have "mixed enough" to explore their entire domain. This is particularly relevant in the case of chains issued from an MCMC algorithm with a multimodal stationary probability, which actually appear in practical situations and for which usual convergence control methods based on graphical evaluations of cumulative sums or similar quantities do not reveal multimodality (see Robert 1996). It appears that normality is reached only when the parallel chains have "spent enough time near every mode" (see the example in Section 5.2), and that multimodal situations are revealed by the occurrence of strong non-normalities for small to moderate values of $n$.

In this section, we investigate the case of finite state Markov chains. The proposed convergence assessments are basically derived from the discrete setting given in §2.2, and rely on the asymptotic behavior of $\sigma_n^2(h)$ and $S_n(\bar{h})/\sqrt{n}$ for $h = \mathbb{I}_i$, $i \in E$, or more generally $h = \mathbb{I}_A$, $A \subset E$. If stationarity is reached for $(x^{(t)})$, then the variance after $n$ steps, $\sigma_n^2(h)$, should be close to the limiting variance $\sigma^2(h)$ and the distribution of $S_n(h)/\sqrt{n}$ should be approximately normal. We propose two complementary methods to guarantee that the CLT can effectively be used after $n$ steps of the algorithm under consideration, in order to build reliable confidence intervals for a class of normalized sums $S_n(h)/\sqrt{n}$. The first method is based on normality assessment and the second one monitors variance stabilization. Both methods use independent parallel chains started from a suitably dispersed distribution $\mu_0$. (See the debate in Gelman and Rubin (1992), and Geyer (1992), about the feasibility of this requirement.)

### 3.1 Convergence assessment by normality monitoring

Basically, the normality control method consists in running $m$ parallel chains $x_1^{(t)}, \ldots, x_m^{(t)}$ started from some preassigned distribution, and testing a normality hypothesis $H_0$ for the r.v.'s $N_n(i)$, $i \in E$, at arbitrary selected times $n_1 < n_2 < \cdots$, until acceptance of normality. It is important noting that using a normality test at successive times here is *only* a way of avoiding graphical monitoring of the approximate normality. It should not be understood as a manner to test the normal model as in the usual statistical practice (with a frequentist interpretation of the type I risk). Consider first a single state $i \in E$. Define

$$N_n^{(\ell)}(i) = \sum_{t=1}^{n} \mathbb{I}\left(x_\ell^{(t)} = i\right), \quad 1 \leq \ell \leq m,$$

the occupation time of state $i$ for chain $\ell$ during the first $n$ steps. We propose to check approximate normality using the Shapiro-Wilk test (Shapiro and Wilk 1965) with a preassigned significance level $\alpha$ to be tuned. This test is one of the most powerful tests against alternative hypotheses as general as "the sample is issued from a non-normal continuous distribution". The Shapiro-Wilk test statistic $SW$ belongs to $(0,1)$, and assumes values close to 1 if the null hypothesis $H_0$ is true (see, e.g., Capéraà and Van Cutsem 1988). It does not require prior knowledge of the expectation $n\pi_i$ of $N_n(i)$. For $m$ chains with initial distribution $\mu_0$, and arbitrary increasing times $n_0 = 0 < n_1 < n_2 < \cdots$, the control method starts with $k = 1$ and proceeds as follows:

1. `Run the` $m$ `chains for` $(n_k - n_{k-1})$ `more iterations.`

2. `Update the sample` $\left(\dfrac{N_{n_k}^{(1)}(i)}{\sqrt{n_k}}, \ldots, \dfrac{N_{n_k}^{(m)}(i)}{\sqrt{n_k}}\right)$. [1]

3. `Compute the Shapiro-Wilk statistic` $SW(i, n_k)$.
   `If` $H_0$ `is rejected,`
   `set` $k \leftarrow k+1$ `and go to 1;`
   `else return` $n_k$.

Let $A_{H_0,\alpha}$ be the acceptance region associated with the level $\alpha$; this algorithm returns $\mathcal{T}_i = \inf_{k \geq 1}\{n_k : SW(i, n_k) \in A_{H_0,\alpha}\}$, the first time in the sequence of $n_k$'s for which the hypothesis has not been rejected. We may in addition plot $SW(i, \cdot)$ and monitor its stabilization in $A_{H_0,\alpha}$ (see §5). Note that [1] is *not* a sequential test in its classical meaning, since we are not doing hypothesis testing at times $n_1 < n_2 < \cdots$ based on $n_1, n_2, \ldots$ iid observations, but rather testing $H_0$ using a sample of constant size $m$, of iid observations from a distribution depending on $n_1, n_2, \ldots$.

In practice, we need to assess normality of the r.v.'s $N_n(i)$ for states in a subset $E' \subset E$, and Steps 2 and 3 of algorithm [1] are easily modified to simultaneously test all states $i \in E'$ over the same $m$ simulated sequences. The normality control method then returns

$(\mathcal{T}_i, i \in E')$. Automated diagnostics resulting from empirical stopping rules (without graphical monitoring) can be proposed. For instance, a simple rule is to run the chains for at least $\mathcal{T}_M = \max\{\mathcal{T}_i, i \in E'\}$ iterations. Alternative stopping rules can be considered as well. For example, a more conservative rule can be: *"stop at the first time all the controlled states* simultaneously *do not reject the null hypothesis"*. We will call this stopping time $\mathcal{T}_S$ for future reference.

In our use of a normality test, the choice of the individual level $\alpha$ is not a crucial matter, but merely a way of detecting a reasonable stabilization of the test statistic. Clearly, our stopping rules are becoming more and more conservative as $\alpha$ increases (e.g. $\mathcal{T}_M$ increases with $\alpha$), and too large values like $\alpha \geq 10\%$ are not advised (see §5.2).

Finally, the choice of $E'$ (and $K'$) is crucial here, and obviously depends to some extent on the size $K$ of the state space. Whereas controlling the normality of occupation time for a number of randomly selected states can be a good choice for strongly mixing chains, this is obviously not true in multimodal situations. For example, consider a chain with a partition $(E_1, E_2)$ of $E$, corresponding to two modes of $\pi$ such that transitions between $E_1$ and $E_2$ occur with small probabilities. Checking normality for $i \in E_1$, say, and too small values of $n$ would obviously result in strongly multimodal histograms, typically with one mode corresponding to the chains started within $E_1$ (many visits to $i$), and another mode corresponding to the chains started within $E_2$ (no excursion to $E_1$ and no visit to $i$). However, such multimodal histograms would be observed only if several parallel chains started within $E_2$ (i.e. for $\mu_0$ dispersed enough). In such a case, excursions between $E_1$ and $E_2$ for the $m$ chains would be more and more likely to occur as $n$ increases, finally allowing us to accept the normality hypothesis for $N_i(n)$. An ideal choice would select one state near each mode of $\pi$ here. However, this would require a preliminary rough knowledge of the global shape of $\pi$, which is not available in general (also, this requirement is not in the spirit of a generic method).

We will generalize [1] in Section 4 by proposing a more definitive and automated solution, which encompasses both the discrete case with large $K$ and the general (continuous state space) situation. We will then discuss more thoroughly the impact of the tuning parameters.

## 3.2 Convergence assessment by variance comparison

A convergence control tool naturally coupled with the normality monitoring consists in checking whether an estimate of the variance after $n$ steps, $\sigma_n^2(h)$, is close to an estimate of the limiting variance $\sigma^2(h)$. For a single state $i \in E$ (i.e. for $h \equiv \mathbb{I}_i$), the natural estimate of $\sigma_n^2(h)$ based on $m$ parallel chains observed up to the $n$th transition is simply the sample empirical variance over the $m$ chains,

$$\hat{\sigma}_n^2(m,h) = \frac{1}{nm} \sum_{\ell=1}^{m} \left( N_n^{(\ell)}(i) - \bar{N}_n(i) \right)^2, \quad \text{where } \bar{N}_n(i) = \frac{1}{m} \sum_{\ell=1}^{m} N_n^{(\ell)}(i).$$

Besides, an estimate of $\sigma^2(h)$ is available by replacing in (9), (11) and (12), the unknown $\mathbb{P}$, $\mathbb{Z}$, $C$ and $\pi$ with consistent estimates based on the $nm$ available simulated steps. A natural

estimate for the $(j,k)$-th entry of $\mathbb{P}$ is then given by

$$\hat{\mathbb{P}}_{jk}(m,n) = \frac{\frac{1}{m}\sum_{\ell=1}^{m}\sum_{t=1}^{n-1}\mathbb{I}\left(x_\ell^{(t)}=j, x_\ell^{(t+1)}=k\right)}{\frac{1}{m}\sum_{\ell=1}^{m}\sum_{t=1}^{n-1}\mathbb{I}\left(x_\ell^{(t)}=j\right)}. \tag{16}$$

A related estimate of $\pi$ can be obtained from the empirical mean occupation times after $nm$ steps,

$$\hat{\pi}_i(m,n) = \frac{1}{m}\sum_{\ell=1}^{m}\sum_{t=1}^{n}\frac{\mathbb{I}\left(x_\ell^{(t)}=i\right)}{n} = \frac{\bar{N}_n(i)}{n}, \quad 1 \le i \le I. \tag{17}$$

Then

$$\hat{\mathbb{Z}}(m,n) = \left\{I - [\hat{\mathbb{P}}(m,n) - \hat{A}(m,n)]\right\}^{-1}, \tag{18}$$

(11) gives $\hat{C}(m,n)$, and $\hat{\sigma}^2(m,n,h) = h^T\hat{C}(m,n)h$. We are interested here in the asymptotic properties of these estimators when $n$ is fixed and $m$ goes to infinity.

**Proposition 3** *For any initial distribution $\mu_0$ and any fixed integer $n$ large enough, we have, a.s. as $m \to +\infty$:*

*(i)* $\hat{\mathbb{P}}_{jk}(m,n) \to \mathbb{P}_{jk}$,

*(ii)* $\hat{\sigma}^2(m,n,h) \to \sigma^2(h)$,

*(iii)* $\hat{\sigma}_n^2(m,h) \to \sigma_n^2(h)$,

*Proof.* These results follow from the strong law of large numbers, with $n$ large enough, typically to allow any state to be reached from any initial state in less than $n$ steps. To illustrate this, consider the strongly aperiodic case, for which $\mathbb{P}_{jk} > 0$ for any $j$ and $k$. Then running $m \to \infty$ chains for just $n = 1$ step is sufficient for the consistency of $\hat{\mathbb{P}}_{jk}(m,n)$. Generally, *(i)* is proved for fixed $n \ge 2$ since

$$\frac{1}{m}\sum_{\ell=1}^{m}\sum_{t=1}^{n-1}\mathbb{I}\left(x_\ell^{(t)}=j, x_\ell^{(t+1)}=k\right) \to \sum_{t=1}^{n-1}\mathbb{P}_{\mu_0}\left[x^{(t)}=j, x^{(t+1)}=k\right]$$

a.s. as $m \to \infty$, and

$$\begin{aligned}\sum_{t=1}^{n-1}\mathbb{P}_{\mu_0}\left[x^{(t)}=j, x^{(t+1)}=k\right] &= \sum_{t=1}^{n-1}\mathbb{P}\left[x^{(t+1)}=k \mid x^{(t)}=j\right]\mathbb{P}_{\mu_0}\left[x^{(t)}=j\right] \\ &= \mathbb{P}_{jk}\sum_{t=1}^{n-1}\mathbb{P}_{\mu_0}\left[x^{(t)}=j\right].\end{aligned}$$

Similarly,

$$\frac{1}{m}\sum_{\ell=1}^{m}\sum_{t=1}^{n-1}\mathbb{I}\left(x_\ell^{(t)}=j\right)\underset{m\to\infty}{\overset{\text{a.s.}}{\to}}\sum_{t=1}^{n-1}\mathbb{P}_{\mu_0}\left[x^{(t)}=j\right],$$

thus $\hat{\mathbb{P}}_{jk}(m,n)\to\mathbb{P}_{jk}$ a.s. as $m\to\infty$. Using (17) we also have that

$$\hat{\pi}_i(m,n)=\frac{1}{n}\left(\frac{1}{m}\sum_{\ell=1}^{m}N_n^{(\ell)}(i)\right)\underset{m\to\infty}{\overset{\text{a.s.}}{\to}}\pi_i,\quad 1\le i\le I,$$

hence $\hat{\pi}(m,n)\to\pi$ a.s. as $m\to\infty$. Using (18) and (11) gives *(ii)*. The consistency of $\hat{\sigma}_n^2(m,h)$ follows from the strong law of large numbers applied to the iid random variables $\left(N_n^{(\ell)}(i)\right)^2$, $1\le\ell\le m$. □

The control by variance comparison uses the setup already described for the normality control. Actually, the two methods can be executed simultaneously, and step 2 of algorithm [1] needs just to be augmented to compute the estimates $\hat{\sigma}_{n_k}^2(m,h)$ and $\hat{\sigma}^2(m,n_k,h)$. In addition to the stopping rule $\mathcal{T}_M$ issued from the normality control, we end up with plots of $\hat{\sigma}_n^2(m,h)$ and $\hat{\sigma}^2(m,n,h)$ against $n$ from which we may check the stabilization and approximate coincidence of the two variance estimators. Note that widely available software systems with algebraic capabilities (e.g., `Mathematica` or `Matlab`) can be used to solve the inversion involved in (18) without additional work for the end user.

These control methods need some adaptation when $K$ gets large. The monitoring using variance comparison requires the computation of the limiting variance, which may not be feasible for large dimension matrices. In such cases this side of the method reduces to graphical monitoring of the stabilization of $\hat{\sigma}_n^2(m,h)$, with no guarantee against apparent stabilization far from the limiting variance. This would result in wrong convergence diagnostics and biased confidence intervals. This is the reason why this control method should not be used alone, but rather coupled with the normality monitoring. The latter is less affected by the size of the chains from a computational perspective (no matrix is involved in the computations), but it can lead to a dramatically conservative method for large $K$'s. Moreover, estimating the probabilities $\pi_i$ for the $K$ states $i\in E$ is meaningless and misleading when $K$ is large (just think of continuous state spaces). A reasonable way to overcome this problem is to choose a partition $(A_1,\ldots,A_p)$ of $E$ and to apply the normality control method to the corresponding indicator functions $h_j=\mathbb{I}_{A_j}$, $j=1,\ldots,p$. These adapted versions are leading to the proposed extensions for the continuous state case, and we thus reserve their descriptions for Section 4.

# 4 Extension to continuous state chains

In this section, we consider an ergodic Markov chain $\left(x^{(t)}\right)$ with continuous state space $E$ and invariant probability distribution $\pi$ with density $f$. We suppose in addition that $x^{(t)}$

satisfies conditions ensuring that the CLT applies, i.e. for every $h \in L_2(f)$, there exists $0 \leq \sigma^2(h) < +\infty$ such that

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left( h(x^{(t)}) - \mathbb{E}^f [h] \right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(h)\right). \tag{19}$$

For general state space the CLT applies, for instance, when the Markov chain is geometrically ergodic. The basic ideas are first to extend the previous results from renewal theory to atomic Markov chains (the renewal state $i$ being replaced with an atom $A$) and second, to transform general Markov chains to atomic Markov chains by splitting a small set. Various sets of sufficient conditions for the CLT to apply in the context of general MCMC's (e.g., for Metropolis and Gibbs kernels) have been investigated (Tierney 1994). A comprehensive survey can be found in Robert (1996).

We do not base our extension to the continuous case on renewal theory for atoms or small sets. Since the construction of appropriate small sets generally requires a deep knowledge of the transition kernel or the target density $f$ (see Guihenneuc-Jouyaux and Robert 1998), this would result in strongly problem-specific control techniques (thus not in the spirit of this normality control principle). Rather, we suggest to select a finite collection of measurable subsets $A_r \subset E$, $1 \leq r \leq p$, typically "almost" partitioning $E$, and to check normality and variance stabilization of the normalized sums $S_n(h_r)/\sqrt{n}$, where $h_r = \mathbb{I}_{A_r}$ for $1 \leq r \leq p$. We can then only obtain estimates and confidence intervals for the $\mathbb{P}^f(A_r)$'s; moreover, since an estimate of the limiting variance can no longer be algebraically computed with formulas (9) to (12), another control of the variance stabilization must be carried out.

This approach through a partition of $E$ is theoretically valid in the continuous setup and in particular does not require any Markovian assumption on the $h_r\left(x^{(t)}\right)$'s. Furthermore, it can apply to general processes $(x^{(t)})$, provided that they converge to a unique stationary regime and satisfy a strong law of large numbers and a CLT similar to (19). This is of particular interest since in general marginal sequences issued from multivariate MCMC algorithms are *not* Markov chains. Actually, in multivariate situations, we check the normality of posterior marginals for simplicity. In addition, approximate normality of $S_n(h)/\sqrt{n}$ for other functions $h$ may be checked simultaneously. For instance, we have always tested the approximate normality for $h(x) = x$ (or higher moments) in the illustrative examples in Section 5, since posterior means for the parameters are generally desired in Bayesian setups. In multivariate situations, marginal sequences were controlled in the same way.

## 4.1 An automated normality monitoring

We propose a methodology for monitoring a general MCMC algorithm, which is grounded on the normality control for finite state case, deals with the specificities of the continuous state case, and has several advantages: (i) it does not require prior knowledge of the target pdf $f$ (and consequently we propose a generic computer code for the normality control); (ii) it requires very few tuning parameters; (iii) the correct "guess" for these parameters are

given on-line by the computer program, through a few preliminary short runs on a trial and error basis (i.e. wrong parameter settings are quickly detected).

In classical settings where the support $E$ of $f$ is the real line or an infinite denumerable set, we obviously cannot measure the tails of $f$ over $E$ accurately: tail regions with almost zero probability would require a dramatically large number of iterations to reach approximate normality, without a noticeable improvement in the precision over estimates like (2). This is one specificity of the continuous case, therefore a preliminary requirement is to restrict our investigations to a suitable compact subset $\mathcal{A}$ of $E$, which needs to be chosen large enough so that $\mathbb{P}^f(\mathcal{A})$ is close to one. This choice, without preliminary knowledge of $f$, has to be validated by the estimate $\hat{\mathbb{P}}(\mathcal{A})$ given on-line by the algorithm. The normality hypothesis may be checked for a collection of indicator functions $h_r$ of subsets $A_r$ of equal length or volume, such that $\mathcal{A} = \bigcup_{r=1}^{p} A_r$. The "controlled region" $\mathcal{A}$ and the "sharpness" $p$ of this partition of $\mathcal{A}$ are preliminary parameters of the procedure (the sharpness $p$ is directly related to the final desired precision for the approximate picture of $f$ given by the histogram $\big(\hat{\mathbb{P}}(A_1), \ldots, \hat{\mathbb{P}}(A_p)\big)$). Since $\mathcal{A}$ may be chosen fairly large (larger than the unknown support of $f$), a natural idea is then to perform the normality control only over the normalized sums of indicator functions of the $A_r$'s representing a significant probability, e.g. such that $\hat{\mathbb{P}}(A_r) > \varepsilon$ for a tuning parameter $\varepsilon$, where these estimated probabilities are updated and checked on-line along with the parallel simulations. The regions representing a non significant proportion of the total mass are thus simply discarded from the set of controlled regions. They would typically correspond to tails of $f$, regions between almost disconnected modes, or regions outside the support of $f$.

More formally, let $C(n)$ be the set of indicator functions $h_r = \mathbb{I}_{A_r}$ which correspond to subsets of significant estimated probabilities for which the normality hypothesis has not yet been accepted at time $n$, with $C(0) = \{h_1, \ldots, h_p\}$. The number of functions in $C(n)$ is decreasing since, at time $n$, normalized indicator functions that have reached approximate normality, or which correspond to subsets of too small probabilities are deleted from $C(n)$. A validation of this deletion procedure is given on-line by the algorithm, in terms of the estimated mass of the subset $\mathcal{A}_C \subset \mathcal{A}$ consisting of the *controlled* sets at convergence,

$$\hat{\mathbb{P}}(\mathcal{A}_C) = \sum_{r=1}^{p} \mathbb{I}_C(A_r)\, \hat{\mathbb{P}}(A_r), \tag{20}$$

where $\mathbb{I}_C(A_r) = 1$ if $A_r$ has been controlled and finally accepted, and $\mathbb{I}_C(A_r) = 0$ if $A_r$ has been discarded.

For $m$ parallel chains $x_1^{(t)}, \ldots, x_m^{(t)}$ started from an initial distribution $\mu_0$ uniform over $\mathcal{A}$ (or even over a larger subset of $E$), we define the sum $S_n(h)$ for the chain $\ell$ by

$$S_n^{(\ell)}(h) = \sum_{t=1}^{n} h\big(x_\ell^{(t)}\big), \quad \ell = 1, \ldots, m,$$

and the consistent estimate of $\mathbb{P}(A_r)$ over the parallel chains by

$$\hat{\mathbb{P}}(A_r) = \frac{\bar{S}_n(h_r)}{n}, \quad \text{where } \bar{S}_n(h_r) = \frac{1}{m}\sum_{\ell=1}^{m} S_n^{(\ell)}(h_r). \tag{21}$$

For a choice $(\mathcal{A}, p, \varepsilon)$ of the tuning parameters and given arbitrary increasing times $n_0 = 0 < n_1 < n_2 < \cdots$, the algorithm for controlling the target distribution is (starting with $k = 1$):

1. Run the $m$ chains for $(n_k - n_{k-1})$ more iterations.

2. For $r = 1, \ldots, p$ update the samples

$$\left( \frac{S_{n_k}^{(1)}(h_r)}{\sqrt{n_k}}, \ldots, \frac{S_{n_k}^{(m)}(h_r)}{\sqrt{n_k}} \right) .$$

3. For $r = 1, \ldots, p$ update $\hat{\mathbb{P}}(A_r)$; [2]
   update $C(n_k) = \left\{ h_r \in C(n_{k-1}) : \hat{\mathbb{P}}(A_r) \geq \varepsilon(n_k) \right\}$.

4. For each $h \in C(n_k)$:
   compute the statistics $SW(h, n_k)$;
   if $H_0$ is accepted for $SW(h, n_k)$, $C(n_k) \leftarrow C(n_k) \setminus \{h\}$.

5. If $C(n_k) = \emptyset$, return $n_k$;
   else set $k \leftarrow k + 1$ and go to 1.

The sequence $\varepsilon(n)$ which appears in [2] is just a refinement of the probability threshold $\varepsilon$. Its purpose is to lower the effect of the poor estimations of the $\mathbb{P}^f(A_r)$'s which may occur during the first few iterations of the $m$ chains. Actually, we do not want to wrongly discard an $A_r$ from being controlled, just because of an underestimation of $\mathbb{P}^f(A_r)$. Choosing a sequence $\varepsilon(n)$ increasing smoothly from 0 to $\varepsilon$ may avoid such a behavior.

Algorithm [2] returns a "time required to assess normality" $\mathcal{T}_M$ which corresponds, as in the discrete case, to the smallest number of iterations required to reach approximate normality in all the subsets $A_r$ with significant estimated probability. For each $h_r \in C(0)$ controlled up to acceptance of $H_0$, the time to reach approximate normality is

$$\mathcal{T}_r = \min_{k \geq 1} \{ n_k : SW(h_r, n_k) \in A_{H_0,\alpha} \}$$

and $\mathcal{T}_M$ is the maximum of these times. To be meaningful, this result must be validated by the estimated mass of the region on which the control has been imposed, $\hat{\mathbb{P}}(\mathcal{A})$, and the estimated mass of the sets within which approximate normality has been reached, $\hat{\mathbb{P}}(\mathcal{A}_C)$, given by (20). Both probabilities should be close to one, and good settings for $(\mathcal{A}, \varepsilon)$ can be found quickly by trial and error over a few preliminary runs of algorithm [2]. Too small

a value for $\hat{\mathbb{P}}(\mathcal{A})$ indicates a wrong choice for $\mathcal{A}$, which misses a significant proportion of the total mass. Too small a value for $\hat{\mathbb{P}}(\mathcal{A}_C)$ indicates that a significant mass has not been controlled (typically in the tails or between distant modes), and consequently that $\varepsilon$ needs to be lowered down.

The algorithm [2] ends up with the overall normality control time, and a detailed picture of $f$ exhibiting all its specificities (e.g., modal regions), together with precise estimates and confidence intervals for the $\mathbb{P}^f(A_r)$'s, based on reliable normal approximations. Simultaneously, normality control over the mean of the parameter or its higher posterior moments are provided. They may be used to give a more conservative stopping rule, and to compute confidence intervals for the parameter (or marginal coordinates in multivariate situations) using the approximate normality.

## 4.2 Variance comparison

In the continuous setup, we can still consistently estimate the variance after $n$ steps, $\sigma_n^2(h)$, by the sample empirical variance

$$\hat{\sigma}_n^2(m,h) = \frac{1}{nm}\sum_{\ell=1}^{m}\left(S_n^{(\ell)}(h) - \bar{S}_n(h)\right)^2, \tag{22}$$

where $\bar{S}_n(h)$ is given in (21). Unfortunately, the algebraic computations leading to an estimate for $\sigma^2(h)$ are no longer feasible and we need some sort of discretization of the continuous Markov chain to mimic the discrete case. Since we do want to keep the generic aspect of this methodology intact, we are not deriving our discretization from small sets as in Guihenneuc-Jouyaux and Robert (1998). Instead, we propose to apply a discretization directly over the partition $(A_1, \ldots, A_p, A_{p+1})$ of $E$ (where $A_{p+1} = E \setminus A$), by considering the process

$$\xi^{(t)} = \sum_{r=1}^{p+1} r\mathbb{I}_{A_r}(x^{(t)}) \tag{23}$$

which takes values in $\{1, \ldots, p+1\}$. This can be seen as a generalization of the binary discretization proposed by Raftery and Lewis (1992). It is no more valid from a theoretical perspective, since $(\xi^{(t)})$ is not a chain for two reasons: (i) when $(x^{(t)})$ is a Markov chain, $(\xi^{(t)})$ does not (usually) satisfy the *lumpability* condition, and (ii) in multivariate situations, the marginal $(x^{(t)})$ is not even a Markov chain. We could determine a batch size (number of iterations ignored between two recordings of the Markov chain) as in Raftery and Lewis (see §5.1), but this seems to make the implementation more difficult without bringing actual improvement. We thus propose this approximation as a trade-off between theoretically-valid discretization and easily implementable and generic control method.

If we consider the process $\xi^{(t)}$ as a discrete Markov chain over the state space $\{1, \ldots, p+1\}$, with pseudo-transition matrix $\mathbb{P}_\xi$ and related matrices $\mathbb{Z}_\xi$, $A_\xi$ and $C_\xi$ as in §3.2, we can algebraically derive, for each $r \in \{1, \ldots, p\}$, an estimate $\hat{\sigma}_\xi^2(m, n, \mathbb{I}_r)$ of the limiting variance $\sigma_\xi^2(\mathbb{I}_r)$ which can be compared with the estimate of $\sigma_n^2(h_r)$ given by (22) for $h_r = \mathbb{I}_{A_r}$,

which is computed on the continuous chain $(x^{(t)})$. As in the discrete case, this estimate provides a graphical tool to check whether the variance after $n$ steps stabilizes around a value which may be considered here as an approximation of the true limiting variance. From the implementation point of view, [2] needs only to be augmented to compute the above estimates along with the tests for the normality hypothesis, for each $h_r \in C(n_k)$ at each step $k$.

# 5 Examples and comparisons

This section contains several examples for both the discrete and continuous (multidimensional) situations. Our purpose is to illustrate the ability of our method to detect delicate situations such as slowly mixing chains, to show its automated and generic aspects, and to compare it with one of the most popular competing method: the binary control of Raftery and Lewis which is first briefly described below.

## 5.1 Raftery and Lewis' binary approximation

The binary control (Raftery and Lewis 1992, 1996) proposes to approximate the time $t_0$ required to reach convergence, and the sample size $T$ necessary to evaluate (1) with a desired precision. The authors consider the two-state process $z^{(t)}$ derived from $x^{(t)}$ by $z^{(t)} = \mathbb{I}_{x^{(t)} \leq \xi}$, where the threshold $\xi$ is an arbitrary point in $E$. They assume that $z^{(t)}$ is a Markov chain (which is in general wrong) with transition

$$\mathbb{P} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix},$$

known stationary probability $\tilde{\pi}$ and second eigenvalue $\lambda_2$. The warm up time $t_0$ is derived from the condition $\max_{i,j} \left| \mathbb{P}(z^{(t_0)} = i | z^{(1)} = j) - \tilde{\pi}_i \right| \leq \varepsilon_1$ which leads to

$$t_0 \geq \log\left( \frac{\varepsilon_1(\alpha+\beta)}{\alpha \vee \beta} \right) \Big/ \log(|\lambda_2|).$$

The sample size $T$ required for the convergence of $\bar{z}_T = \sum_{t=t_0+1}^{T} z^{(t)}/T$ to $\tilde{\pi}_1$ comes from the condition $\mathbb{P}(|\bar{z}_T - \tilde{\pi}_1| < \varepsilon_2) \geq 1 - \varepsilon_3$, with a normal approximation of $\bar{z}_T$ from which they obtain

$$T \geq \frac{\alpha\beta(\lambda_2+1)}{\varepsilon_2^2(\alpha+\beta)^3} \left[ \Phi^{-1}\left( \frac{2-\varepsilon_3}{2} \right) \right]^2.$$

Since $z^{(t)}$ is not a Markov chain (it does not, in general, satisfy the lumpability condition of Kemeny and Snell (1960)), Raftery and Lewis propose to use a *batch size* $B$ (number of iterations ignored between to recordings of the Markov chain) to approximate independence. However, their procedure for determining $B$ seems rather weak theoretically and outputs quite small values for $B$ (often $B = 1$) in the examples (see Robert 1996). The binary control

only requires a preliminary run to estimate $(\alpha, \beta)$ from which $t_0$ and $T$ are computed. The question of the duration $T_m$ of this preliminary run should itself address a convergence control problem.

The comparison between our approach with parallel chains, and the binary control which uses a single chain needs to be clarified. In particular, we cannot directly compare the precision of the confidence intervals (CIs) for the estimates, since the binary control does not directly deliver CIs for the parameters or the $\pi_i$'s in the discrete case, but merely for $\tilde{\pi}_1$ (and without checking for normality before using the normal approximation). Since both methods produce automated stopping rules, we will do the comparisons in terms of these rules. The binary control proposes to run a single chain for $t_0$ iterations (after what stationarity is supposedly achieved), and then to use the next $T$ jumps to compute the desired estimates. The normality control concludes that $\mathcal{T}_M$ (or $\mathcal{T}_S$) iterations are needed for the normalized sums of the chain to reach normality, but it runs $m$ parallel chains to establish its diagnostic, and then uses all the available information — i.e. the $m \times \mathcal{T}_M$ jumps — to compute estimates. Hence we believe that a fair way to compare the methods consists in facing $t_0$ and $\mathcal{T}_M$ for the burn-in duration, and $T$ against $m \times \mathcal{T}_M$ for the estimation duration. The latter comparison may thus provide answers to the usual question about "single versus parallel chains". We will also compare the true $\pi$ (when available) with the stationary probabilities estimated by following each methods' diagnostics.

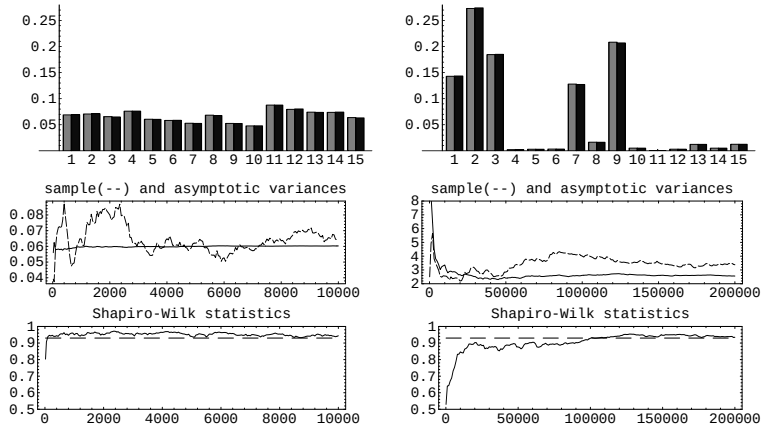
## 5.2 A miniature example

Our purpose here is to illustrate on a toy (but meaningful) example for $K = 15$ states the control methods of Section 3, and to show how the stabilization processes of the variance estimates and test statistics may differ between two finite chains. We defined a chain $\left(x^{(t)}\right)$ with a random transition matrix allowing for quick transitions between all states, and having consequently a roughly uniform invariant probability $\pi(x)$ (Figure 1, *top left*). We constructed also a chain $\left(y^{(t)}\right)$ with a transition matrix tailor-made for generating a multimodal invariant probability $\pi(y)$ with three modal regions $E_1 = (1, 2, 3)$, $E_2 = (7, 8, 9)$ and the smaller mode $E_3 = (13, 14, 15)$ (Figure 1, *top right*). The transitions were chosen so that jumps between modal regions follow the scheme $E_1 \leftrightarrow E_2 \leftrightarrow E_3 \rightarrow E_1$, resulting in a slowly mixing chain.

For both examples, we ran $m = 50$ independent parallel chains, started according to a uniform initial distribution over $E$. We then controlled four to six states with algorithm [1] together with the variance comparison of §3.2, using the asymptotic level $\alpha = 0.01$. We implemented the two stopping rules $\mathcal{T}_M$ and (the more conservative) $\mathcal{T}_S$ defined in §3.1. The estimated invariant probabilities in Figure 1, and Student's $t$ 95%–confidence intervals (CIs) based on the normality assumption for the sample of occupation times were computed at the former stopping time, which was similar to $\mathcal{T}_S$ in this example. We always ran the parallel chains up to a fixed large amount of iterations to show stabilization after the stopping time.

For the simple chain $x$, convergence occurred after only 100 to 200 iterations, whatever the four arbitrarily selected controlled states, and for both stopping rules. All the CIs at this time contained the true values. Figure 1 (*bottom left*) shows a typical output for

Figure 1: *Top*: Invariant probabilities $\pi(x)$ *(left)* and $\pi(y)$ *(right)*. The true probabilities are in gray, and their estimates at convergence time in black. *Bottom, left*: chain $x$, control for state 2. The true variance is $\sigma^2(\mathbb{I}_2) = 0.0598$. The stopping rule gives $\mathcal{T}_2 = 100$. *Bottom, right*: chain $y$, control for state 15. The true variance is $\sigma^2(\mathbb{I}_{15}) = 2.494$. The stopping rule gives $\mathcal{T}_{15} = 101,000$.




state 2 for which we obtain the 95%-CI $[0.060, 0.074]$ for the true value $\pi_2(x) = 0.0706$ after 200 iterations. Since $\mathcal{T}_M$ is small, we just ran the chains up to 10,000 iterations. The stabilization of $\hat{\sigma}_n^2(m, \mathbb{I}_2)$ seems to be achieved after about 2400 iterations, and shows merely noise around $\hat{\sigma}^2(m, n, \mathbb{I}_2)$ after that. The latter estimate stabilizes in a time comparable to $\mathcal{T}_M$. The heuristic procedure of Section 2.3 using the Berry-Esséen inequality (14) required a dramatically too large convergence time of about $10^{15}$ iterations for this example.

For the "multimodal" chain $y$, we applied our controls over one state near each mode and one state between two contiguous modal regions. Here, convergence (in the normality control sense and for both stopping rules) required not less than 100,000 iterations. Again, at convergence, the Student's $t$ 95%–CIs contained the true values for each of the controlled states, and we ran the chains up to 200,000 iterations to show stabilization. For states in the most frequently visited regions $E_1$ and $E_2$ stabilization was achieved comparatively quickly (about 10,000 iterations). Surprisingly, the same behavior held for state 5 between $E_1$ and $E_2$, although $\pi_5(y) = 0.00298$ is very small. Finally, the normality check provided the most conservative time for states in $E_3$ like, e.g., state 15 displayed here (Figure 1, *bottom, right*). Although $\pi_{15}(y)$ is four times larger than $\pi_5(y)$, the normality control required 101,000 iterations, and the Shapiro-Wilk statistic stabilized after that. Note that the procedure based on the Berry-Esséen inequality (14) required up to $10^{25}$ iterations to assess convergence.

This typical behavior for the chain $y$ arises because many jumps occur between the first two modes. Therefore, the chain frequently visits states $4, 5, 6$, leading rather quickly to a

normal-like histogram for those states. When started outside $E_3$, a chain usually does not visit it for a very long time. Roughly 1/5 of the 50 chains started from this third mode and spent time within it, whereas the remaining 4/5 started outside this mode and remained outside it for many iterations. This explains why samples of occupation times for state 15 are far from normality at the beginning. In this example, about 100,000 iterations allow for enough visits to $E_3$ for most of the 50 chains to attain normality. This miniature example shows that it is basically multimodality, and not only estimation of small probabilities for the stationary distribution (e.g., $\pi_5(y)$), that really affects the normality of samples of occupation times. It also highlights the ability of our parallel chain method, started with a dispersed initial distribution, to detect such delicate situations.

As pointed out in Section 3.1, the choice of $\alpha$ is not crucial. The $\alpha$-quantile of the Shapiro-Wilk statistic (the horizontal dashed lines in Figure 1, *bottom*) increases with $\alpha$, resulting in more conservative rules. Reasonable modifications of $\alpha$ give stopping times of same order. For instance, the normality control with $\alpha = 5\%$ delivered here $\mathcal{T}_M = 700$ for the chain $x$, and $\mathcal{T}_M = 123,000$ for the chain $y$. However, too large values of $\alpha$ are not recommended, e.g., the 10%-quantile was never reached during the first $400,000$ iterations of chain $y$.

**Comparison with the binary control**

For each chain $x$ and $y$, we ran the binary control with starting values drawn uniformly in $E$, fixed errors $\varepsilon_1 = 0.001$ and $\varepsilon_2 = \varepsilon_3 = 0.01$, batch sizes $B = 1, 2, 10, 20$ and pre-run sizes $T_m = 1000$, $10,000$ and $100,000$. The binary control is extremely sensitive to the choice of $\xi$, and we illustrate this by running the control for each $\xi \in E$ and summarizing the results using the most conservative stopping rules $\max_{\xi \in E} t_0(\xi)$ and $\max_{\xi \in E} T(\xi)$, and the less conservative ones $\min_{\xi \in E} t_0(\xi)$ and $\min_{\xi \in E} T(\xi)$.

For the quickly mixing chain $x$, results were not significantly influenced by our choices for the parameters $(T_m, B)$, and we obtained $1 \leq t_0 \leq 3$, and $3909 \leq T \leq 17,616$. By comparison, the normality control gives $100 \leq \mathcal{T}_M \leq 200$ and $5000 \leq m \times \mathcal{T}_M \leq 10,000$. The values proposed by the binary control for $t_0$ are indeed not realistic and $\mathcal{T}_M$ should be preferred, but both methods compute their estimates based on a number of iterations of the same order.

For the slowly mixing chain $y$, results are displayed in Table 1 ($B = 20$ provided results similar to $B = 10$; $(T_m = 1000; B = 1)$ is not given since $(\alpha, \beta)$ could not be computed with that setting for most of the $\xi$'s). The incredibly wide range of variation in the rules given by the binary control, depending on $\xi$ but also on $T_m$ and $B$, shows that this method should be used with caution. By comparison, the normality control gives about $\mathcal{T}_M = 100,000$ and then computes the estimated $\pi$ using $m \times \mathcal{T}_M = 5,000,000$ iterations. Note that even a preliminary run of $T_m = 100,000$ iterations for computing $\alpha$ and $\beta$ does not improve the binary control's results.

To find which method provides the best guess, we compared the qualities of the resulting estimates of $\pi$. For the binary control, we started a single chain from a value drawn uniformly in $E$, ran $t_0$ iterations, and used the subsequent $T$ iterations to compute the $\hat{\pi}_i$'s from the

Table 1: Binary control results for the chain $y$ and selected tuning parameters.

| $T_m$ | $B$ | $\min(t0)$ | $\max(t0)$ | $\min(T)$ | $\max(T)$ |
|---|---|---|---|---|---|
| 1,000 | 2 | 2 | 4,653 | 1 | 2,834,980 |
| 1,000 | 10 | 1 | 966 | 1 | 3,484,659 |
| 10,000 | 1 | 1 | 9,676 | 7 | 34,887,061 |
| 10,000 | 2 | 4 | 5,972 | 4,249 | 21,511,126 |
| 10,000 | 10 | 4 | 800 | 692 | 4,188,248 |
| 100,000 | 1 | 2 | 8,029 | 691 | 42,028,953 |
| 100,000 | 2 | 4 | 5,304 | 1,851 | 25,924,892 |
| 100,000 | 10 | 4 | 971 | 1,856 | 4,653,923 |

Table 2: Chi-square distance between the true and estimated $\pi(y)$ for selected couples $(t_0, T)$ reflecting typically less and more conservative rules. The last row gives the normality control, based on $m\mathcal{T}_M = 5$ millions iterations.

| $T_m$ | $B$ | $t_0$ | $T$ | $\chi^2(\pi(y), \hat{\pi}(y))$ |
|---|---|---|---|---|
| 1,000 | 2 | 3 | 1,091 | 0,654445 |
| 1,000 | 2 | 979 | 2,834,980 | 0,000242 |
| 10,000 | 2 | 9 | 4,249 | 0,104466 |
| 10,000 | 2 | 5,972 | 21,511,126 | 0,000188 |
| 100,000 | 2 | 8 | 1,851 | 0,655096 |
| 100,000 | 2 | 5,304 | 25,924,892 | 0,000158 |
| NC | | 100,000 | 5,000,000 | 0,000024 |

observed occupation times. For the normality control, the equivalent estimates are the $\hat{\pi}_i(m, n)$'s given by (17). Table 2 shows the results for $y$, in terms of the chi-square distance, for selected solutions $(t_0, T)$ proposed by the binary control, and for the normality control. It appears that many choices for $\xi$ would lead to dramatically wrong estimates for $\pi$ with the binary control. Still more impressive, even the most conservative rule given by the binary control results in a $\hat{\pi}$ which is not as good as the estimate given by the normality control. The latter estimate is computed with far less iterations than the former though (5 millions against 25 millions), but it uses a parallel method. We observed similar results for the simplest chain $x$, for which the binary control with a conservative rule gives $\chi^2(\pi(x), \hat{\pi}(x)) = 0.00054$, and the normality control gives $\chi^2(\pi(x), \hat{\pi}(x)) = 0.000036$. This illustrates the benefits of our approach for slowly mixing chains, and the advantages of parallel chains versus a single chain.

## 5.3 Comparison between stationarity and normality

Relations between stationarity and normality are not clear, at least from a mathematical point of view. There is no obvious theoretical reason to think that one is reached before or

after the other, but if we are interested in confidence intervals, we have to collect enough information relative to occupation times to construct confidence intervals based on the normality assumption. Furthermore, this would still hold true even if we knew the stationary distribution and actually used it as the initial distribution. In order to compare the times needed to reach stationarity and normality, respectively, we selected the discrete example below, for which the number of iterations needed to reach approximate stationarity can be theoretically computed.

We considered the random walk $X$ on the $d$-dimensional cube $E = \{0,1\}^d$. Two states in $E$ are connected together (neighbors) if they differ only by exactly one coordinate, i.e. if $\sum_{i=1}^{d} |x_i - y_i| = 1$. For $0 < \beta < 1$, the transition matrix of this chain is given by:

$$\begin{cases} P_{x,x} = 1 - \beta, \quad x \in E \\ P_{x,y} = \beta/d \quad \text{if } x \text{ is connected to } y \\ P_{x,y} = 0 \quad \text{if } x \text{ is not connected to } y. \end{cases}$$

This Markov chain has been studied in details by, e.g., Diaconis, Graham and Morrison (1990). In particular, when $\beta \leq d/(d+1)$, the second eigenvalue of the transition matrix can be analytically computed and for each small $\varepsilon > 0$, the time to reach $\varepsilon$-approximate stationarity (in the total variation norm) can be expressed by the following relation:

$$\text{if} \quad n > \frac{d}{4\beta}\big(\log d - \log\log(1+\varepsilon)\big) \quad \text{then} \quad ||\mathbb{P}_{X_n} - \pi||_{TV} < \varepsilon, \tag{24}$$

where $\mathbb{P}_{X_n}$ denotes the distribution of $X_n$. To illustrate relations between stationarity and normality in this case, we simulated this random walk for dimensions $d = 3$ and $d = 4$, and several values of $\beta$ decreasing from 0.5 to 0.01. We always run 50 parallel chains, and applied the normality control over 8 states, that is *all* the state space for $d = 3$ and half of it for $d = 4$ (states were chosen randomly in the latter case). For simplicity, we still took $\alpha = 0.01$. For each dimension, we plotted the following number of iterations against $\beta$ values:

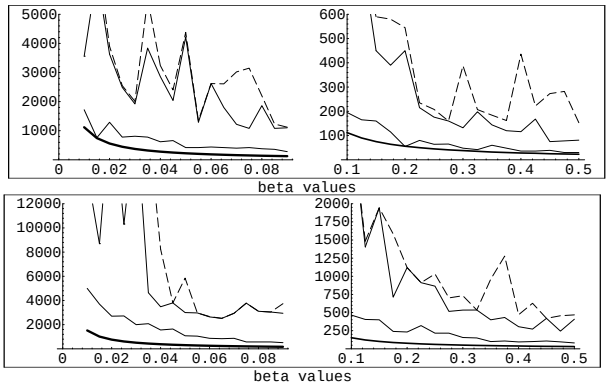
$\mathcal{T}_\varepsilon$ the theoretical time required to attain an $\varepsilon$-stationary regime for $\varepsilon = 10^{-6}$, as given by (24) (solid bold lines in Figure 2);

$\mathcal{T}_{\min}$ the time of first entrance into the acceptance domain, i.e. the minimum over the controlled states of each state's first entrance time (solid line above the previous one);

$\mathcal{T}_M$ the time of last entrance into the acceptance domain (i.e. the time given by our first stopping rule, solid line above the previous one);

$\mathcal{T}_S$ the conservative stopping rule for the normality assessment, as defined in §3.1, i.e. the first time at which all the controlled states simultaneously accept the null hypothesis (dashed line).

Figure 2: Measures of convergence for the random walk with $d = 3$ (*top*) and $d = 4$ (*bottom*). The represented convergence times are, from top to bottom: $\mathcal{T}_S$ (dashed), $\mathcal{T}_M$, $\mathcal{T}_{\min}$ and $\mathcal{T}_\varepsilon$ (bold).




These times satisfy $\mathcal{T}_{\min} \leq \mathcal{T}_M \leq \mathcal{T}_S$. Figure 2 displays these measures of convergence expressed in numbers of iterations for $\beta \in [0.1, 0.5]$ and $\beta \in [0.01, 0.09]$ (in separate pictures for better readability), and for $d = 3$ and $d = 4$.

In this experiment, $\mathcal{T}_{\min}$ is close to the theoretical value $\mathcal{T}_\varepsilon$, and the three empirical measures are increasing (as $\mathcal{T}_\varepsilon$ does) when $\beta$ decreases to zero (i.e. when the chains mix slower). Moreover, $\mathcal{T}_M$ and $\mathcal{T}_S$ are both really conservative. Reaching approximate normality seems to take more time than reaching approximate stationarity. This seems natural, since even a stationary process needs to run for a while to obtain a useful precision. We should notice that we controlled 8 states, giving obviously a rather conservative procedure, especially for $\mathcal{T}_S$. Finally, the plots for the normality control (state by state), which are not displayed here for brevity, always showed a clear stabilization after the convergence time $\mathcal{T}_M$, with only some short excursions out of the acceptance regions.

We also ran the binary control on this example, with $\varepsilon_1 = 10^{-6}$ i.e. the precision set for $\mathcal{T}_\varepsilon$. The results are less sensitive to $\xi$ than in 5.2, due to the total symmetry in $\xi$ of this example. Results for $d = 3$ and several $\beta$ values are in Table 3, where $T$ is again compared with $m \times \mathcal{T}_M$. The conservative rule for $t_0$ is close to the theoretical time to reach $\varepsilon$-stationarity here, but the proposed conservative rules for $T$ are again overestimated when the chains mix slower, as shown by Table 4 which gives the chi-square distances. The estimates given by the binary control are better than ours for quickly mixing chains, but they are computed using much more iterations. For a slowly mixing random walk (e.g., here for $\beta = 0.01$), our parallel method gives a comparable estimate of $\pi$ in 14 times less iterations.

Table 3: Binary control versus normality control for the random walk ($d = 3$).

| $\beta$ | $\mathcal{T}_\varepsilon$ | $min(t_0)$ | $\max(t_0)$ | $\min(T)$ | $\max(T)$ | $m \times \mathcal{T}_M$ |
|---|---|---|---|---|---|---|
| 0.50 | 23 | 10 | 17 | 12,385 | 43,373 | 4,050 |
| 0.40 | 28 | 13 | 21 | 14,884 | 54,830 | 5,800 |
| 0.30 | 38 | 17 | 30 | 19,046 | 75,606 | 6,600 |
| 0.20 | 56 | 28 | 47 | 30,993 | 117,747 | 22,500 |
| 0.10 | 112 | 58 | 97 | 59,932 | 244,131 | 32,750 |
| 0.09 | 125 | 59 | 109 | 56,656 | 274,765 | 56,000 |
| 0.05 | 224 | 106 | 204 | 102,095 | 513,990 | 143,000 |
| 0.01 | 1,119 | 731 | 1004 | 901,048 | 2,533,891 | 178,000 |

Table 4: Chi-square distance between the true $\pi$ and estimates $\hat{\pi}$ given by the normality control (NC) and the binary control's most conservative rules (BC), for the random walk ($d = 3$).

| $\beta$ | $t_0$ | $T$ | $m \times \mathcal{T}_M$ | $\chi^2(\pi, \hat{\pi}_{NC})$ | $\chi^2(\pi, \hat{\pi}_{BC})$ |
|---|---|---|---|---|---|
| 0.5 | 17 | 43,373 | 4,050 | 0.001331 | 0.000323 |
| 0.1 | 97 | 244,131 | 32,750 | 0.002152 | 0.000701 |
| 0.01 | 1004 | 2,533,891 | 178,000 | 0.000595 | 0.000520 |

## 5.4 A continuous example with multimodal posterior

To illustrate our methodology for continuous state chains, we consider the following example from Robert (1996), which results in a multimodal target density. This example is relative to Bayesian inference for the location parameter $\theta$ of a Cauchy $\mathcal{C}(\theta, 1)$ distribution using a normal prior $\mathcal{N}(0, \sigma^2)$. With three observations $x_1, x_2, x_3$ from $\mathcal{C}(\theta, 1)$, the posterior density is

$$\pi(\theta|x_1, x_2, x_3) \propto \exp\left(\frac{-\theta^2}{2\sigma^2}\right) \left[\prod_{i=1}^{3} \left(1 + (\theta - x_i)^2\right)\right]^{-1} \tag{25}$$

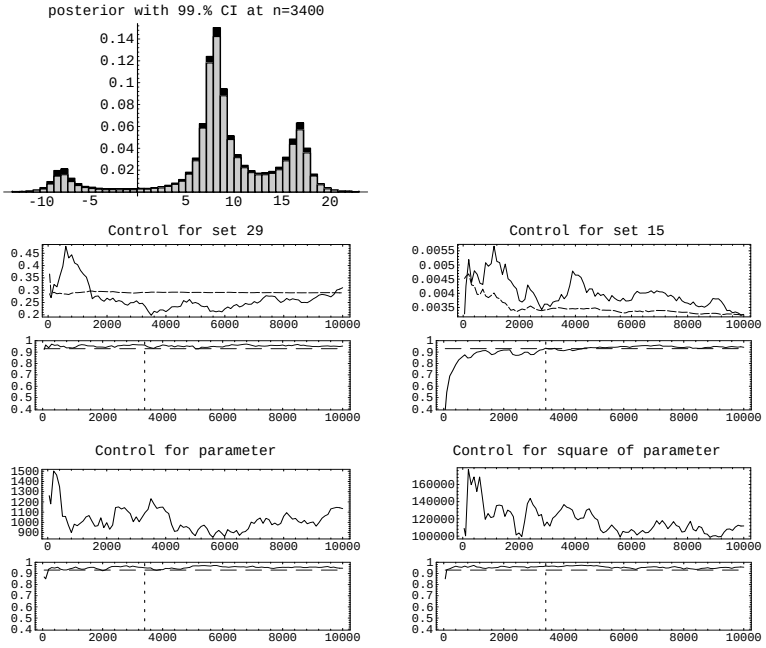
and (25) appears as a marginal of the augmented density

$$\pi(\theta, \eta_1, \eta_2, \eta_3|x_1, x_2, x_3) \propto \exp\left(\frac{-\theta^2}{2\sigma^2}\right) \prod_{i=1}^{3} \exp\left[-\left(1 + (\theta - x_i)^2\right) \eta_i/2\right], \tag{26}$$

with the three instrumental r.v.'s $\eta_1, \eta_2, \eta_3$. The conditional distributions derived from (26) are available for simulation, and Robert (1996) proposes the following Gibbs Sampler for simulation from (25):

1. $(\eta_i|\theta, x_i) \sim \mathcal{E}\left(\frac{1+(\theta-x_i)^2}{2}\right)$, $i = 1, 2, 3$.
2. $(\theta|\eta_1, \eta_2, \eta_3, x_1, x_2, x_3) \sim \mathcal{N}\left(\frac{\sum_{j=1}^{3} \eta_j x_j}{\sum_{j=1}^{3} \eta_j + \sigma^{-2}}, \left(\sum_{j=1}^{3} \eta_j + \sigma^{-2}\right)^{-1}\right)$.

Figure 3: Estimated posterior distribution (25) at convergence time with Student CIs (*in black*). Second row, control for $A_r$'s corresponding to the fastest (*left*) and slowest (*right*) convergence times. Third row, control for $\theta$ and $\theta^2$. Each control consists in two plots: the variance comparison with the approximate asymptotic variance in dashed lines when available (*top*), and the Shapiro-Wilk statistic, with its acceptance region above the dashed line (*bottom*).




For the selected observations $x_1 = -8$, $x_2 = 8$ and $x_3 = 17$, (25) is trimodal with a large gap between the modes located in $-8$ and $+8$.

This example is one-dimensional, hence our automated normality control method of §4.1 directly applies. In order to illustrate the on-line determination of the tuning parameters $(\mathcal{A}, p, \varepsilon)$, we first blindly ran [2] for 1000 iterations of $m = 50$ chains on the erroneous region $\mathcal{A} = [0, 200]$, with $p = 20$ sets and $\varepsilon = 0.004$. Rightmost sets were quickly discarded, and convergence occurred in remaining sets. But the wrong choice for $\mathcal{A}$ was revealed by the poor estimates $\hat{\mathbb{P}}(\mathcal{A}) = 90.7\%$ and $\hat{\mathbb{P}}(\mathcal{A}_C) = 90.6\%$, indicating that some significant part of the mass (the third leftmost mode) is missing. The selection of an appropriate $\mathcal{A}$ was then attained by a few runs on this trial and error basis, and resulted in the estimate $\hat{\mathbb{P}}(\mathcal{A}) = 99.7\%$. We imposed a sharpness $p = 50$ here, to get a reasonable precision for the trimodal histogram of the posterior distribution, and the threshold for controlling the sets $A_r$'s had to be lowered to $\varepsilon = 0.002$ to gain control over $\hat{\mathbb{P}}(\mathcal{A}_C) = 99\%$ of total mass. Note

Table 5: Binary control results with $T_m = 10,000$ and batch sizes $B$ for the multimodal Cauchy example.

| $B$ | $\min(t_0)$ | $\max(t_0)$ | $\min(T)$ | $\max(T)$ |
|---|---|---|---|---|
| 1 | 4 | 60 | 507 | 105,958 |
| 2 | 2 | 35 | 360 | 61,253 |
| 10 | 2 | 10 | 353 | 23,566 |

that $p$ and $\varepsilon$ are linked somehow, since the probabilities $\mathbb{P}^f(A_r)$'s decrease when $p$ increases, and consequently, the times needed to reach approximate normality increase (mostly for the $A_r$'s in the tails or located between modes, where less and less jumps are observed when the $|A_r|$'s decrease). For this reason, $p$ should not be larger than a value imposed by the precision wanted for the "picture" of $f$. In other words, the more precision we want for the histogram of $f$, the more time it takes to get this picture with approximate normality. In addition to our control over the posterior distribution, we controlled the functions $g_1(\theta) = \theta$ and $g_2(\theta) = \theta^2$. Note that the additional stopping rules associated with the control of $g_1$ and $g_2$ are not influenced by the choices made for $p$ and $\varepsilon$, and in this sense act as moderators.

Approximate normality occurred around $n = 3400$ iterations, and we ran the $m = 50$ parallel chains up to $10,000$ iterations to show stabilization. Figure 3 shows some selected results. As expected, the stopping rule $\mathcal{T}_M$ corresponds to a set $A_{15}$ located around $-2$, between the largest mode and the smallest distant mode (Figure 3, *right*). The plots for $g_1$ and $g_2$ show a quick stabilization; normality is reached after less than 500 iterations. This indicates that the region of small probability between the two distant modes, although requiring 3400 iterations to reach approximate normality, has little influence over the estimation of $\theta$ or $\theta^2$. Note that the approximate estimates for the limiting variance $\hat{\sigma}_\xi(m, n, h_r)$, available just for the indicator functions, behave as in the discrete case: they stabilize rather quickly, but not always around the average value of the sample variance. This may be a side-effect of the discretization, or an effect of the long-memory which characterizes the sample variance process. However, as in the discrete case, they may be used as a complementary tool coupled with the normality control.

We ran the binary control for this Gibbs sampler with the settings already used in §5.2, and several values for $\xi \in [-10; 20]$. Less and most conservative rules are summarized in Table 5, and show again rather small values for $t_0$, and a wide range of variation for $T$. The normalization constant of (25) could be computed with `Mathematica` here, and an approximation of the cdf was available. We thus compare the estimates with the chi-square distance between a discretized version of the true density (25) and the similar histograms obtained by following each method's stopping times (as in §5.2). The results in Table 6 show that the normality control gives the most conservative rule. Our method seems to detect the difficulties linked to multimodality, since it also provides the best estimate of the density.

Table 6: $\chi^2$ distances between the discretized $\pi$ of the Cauchy example and estimates using the binary control (BC) and the normality control (NC) rules.

| Method | $t_0$ | $T$ | $\chi^2(\pi, \hat{\pi})$ |
|---|---|---|---|
| BC | 2 | 360 | 0.17834 |
| BC | 10 | 23,566 | 0.00395 |
| BC | 60 | 105,958 | 0.00256 |
| NC | 3,400 | 170,000 | 0.00138 |

## 5.5 Mixture of distributions

We consider here a missing data model with a 5-dimensional parameter, consisting in observations issued from a two-component normal mixture distribution

$$p\mathcal{N}(\mu_1, \sigma_1) + (1-p)\mathcal{N}(\mu_2, \sigma_2) \tag{27}$$

in a Bayesian framework with parameter $\theta = (p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ and conjugate priors

$$p \sim \mathcal{B}e(1/2, 1/2), \quad \mu_i \sim \mathcal{N}(\xi_i, \sigma^2/n_i), \quad \sigma_i^2 \sim \mathcal{IG}(\nu_i/2, \omega_i^2/2), \quad (i = 1, 2).$$

We generated a sample of size 30 from (27) with the true parameter $\theta^* = (0.3, -3, 1, 3, 4)$, and used the Gibbs implementation given in Diebolt *et al.* (1994), which iteratively simulates the missing data and the parameter $\theta$. We then applied the normality control [2] together with the variance comparison marginally for each parameter's posterior distribution, marginally for each scalar coordinate (e.g., for functions $g_i(\theta) = \theta_i$, $i = 1, \ldots, 5$), and also over the scalar function $g(\theta) = p + \mu_1 + \sigma_1^2 + \mu_2 + \sigma_2^2$. The selection of the controlled region $\mathcal{A}$ for each coordinate was easily done by short runs of [2] for a few iterations. It sketched out the mass location for each marginal posterior. The resulting estimates $\hat{\mathbb{P}}(\mathcal{A})$ were always larger than 99.9%. The threshold was set to $\varepsilon = 0.004$, resulting in estimates for $\hat{\mathbb{P}}(\mathcal{A}_C)$ between 98.5% and 99.3%. Convergence in our sense, and for all the controlled functions, always occurred before $n = 2000$ iterations, and we ran the $m = 50$ parallel chains up to 10,000 iterations to show stabilization.

Figure 4 shows a set of results for two coordinates. The estimates $\left(\hat{\mathbb{P}}(A_1), \ldots, \hat{\mathbb{P}}(A_p)\right)$ at time $\mathcal{T}_M$ are represented by the histograms together with their confidence intervals using the normal approximation. We then give for each coordinate control plots for the $h_r = \mathbb{I}_{A_r}$ requiring the largest time to reach approximate normality (typically in the tails here). In this example, the function $g(\theta)$ which can be seen as a global control for this MCMC algorithm, required 1600 iterations to reach approximate normality. Actually, the largest convergence time was obtained for the parameter $\sigma_1^2$ (i.e. for the control over $g_5(\theta) = \sigma_1^2$) which required 2000 iterations. On the other side, the indicator functions of the $A_r$'s located near the modes of the posterior marginals stabilized in less than 100 iterations, as expected.

We ran the binary control marginally on each coordinate, with the settings $\varepsilon_1 = 0.001$ and $\varepsilon_2 = \varepsilon_3 = 0.01$, several batch sizes and pre-run $T_m$ between 1000 and 10,000. Here

Figure 4: Control for the mixture parameters $\mu_1$ (*left*), and $\sigma_2^2$ (*right*). Each column gives successively the estimated marginal posterior distributions at convergence time with student CIs (*in black*), the control for selected $A_r$'s, and the control for the coordinates. The last row is the control for $g(\theta)$. Control plot are as in Figure 3.

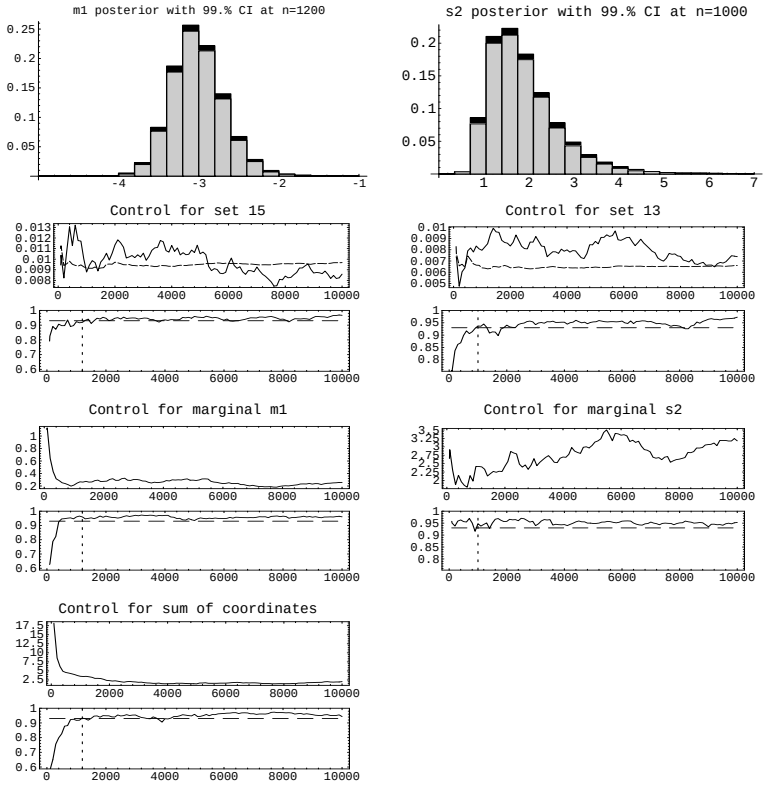


Table 7: Binary control for the mixture model ($B = 2$ and $T_m = 3000$).

| $\theta_i$ | $\max(t_0)$ | $\min(T)$ | $\max(T)$ |
|---|---|---|---|
| $p$ | 2 | 392 | 14,761 |
| $\mu_1$ | 3 | 14 | 20,263 |
| $\sigma_1$ | 5 | 840 | 24,020 |
| $\mu_2$ | 3 | 4,039 | 18,804 |
| $\sigma_2$ | 4 | 567 | 21,852 |

again, the batch size has a limited influence over $(t_0, T)$, but the thresholds $\xi$ really influence $T$, and they have to be selected from rough approximations of the posterior marginals of each scalar parameter, since they must be in the support and not too far in the tails, for the binary control to be worked out. Again, we observed unrealistic values between for $t_0$, and a great variability in the proposed stopping rules for $T$, which are summarized in Table 7. By comparison, the global stopping rule given by the normality control was about $\mathcal{T}_M = 2000$, which is in accordance with the observed stabilization for the test statistics and the empirical variances (see, e.g., Figure 4). In this case, the true posterior is unknown. Therefore we cannot compare the methods on the basis of the quality of their estimates.

# 6 Conclusion

The aim of this paper was to propose a new method for controlling the convergence of MCMC algorithms, based on parallel independent realizations of the Markov chain started from an initial distribution dispersed enough to ensure a thorough and efficient exploration of the support of the target density. Our purpose was to check that the distributions of a finite collection of normalized sums of functions of the Markov chain have reached approximate normality and that their variances have stabilized around their asymptotic value. We could then construct reliable Student-Confidence intervals for the corresponding approximations. Our methodology has several advantages from the end user point of view; it is not problem-specific, i.e. it is completely independent from the MCMC algorithm under consideration and a generic computer code implementing our methods is available. This code provides automated stopping rules which do not require deep experience on the part of the user.

We have reported on numerical investigations and comparisons with the binary control in both finite and continuous settings. These experiments show that: (i) test levels $\alpha$ between 1% and 5% are suitable for most examples, and result in comparable stopping times; (ii) the tuning parameters $(\mathcal{A}, p, \varepsilon)$ are easily determined on-line by trial and error; (iii) slowly mixing chains and multimodal invariant distributions have been satisfactorily recovered; (iv) the binary control provides rules with a large variability, and its extreme stopping times are often not reasonable; (v) comparisons between the estimates given by each method and the true distribution argue for parallel chains against a single chain (better estimates in a smaller total number of iterations).

## References

[1] Billingsley, P. (1986), *Probability and Measure*, 2nd Ed., John Wiley & Sons, New York.

[2] Bolthausen, E. (1982), "The Berry-Esséen Theorem for Strongly Mixing Harris Recurrent Markov Chains," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 60, 283–289.

[3] Brooks, S., and Roberts, G. (1995), "Diagnosing Convergence of Markov Chain Monte Carlo Algorithms", Tech. report 95–12, Stat. Lab., U. of Cambridge.

[4] Capéraà, P., and Van Cutsem, B. (1988), *Méthodes et Modèles en Statistique non paramétrique*, Dunod.

[5] Chung, K.L. (1967), *Markov Chains with Stationary Transition Probabilities*, Springer-Verlag, Berlin.

[6] Dacunha-Castelle, D. and Duflo, M. (1986), *Probability and Statistics*, vol. 2, Springer-Verlag, New York.

[7] Diaconis, P., Graham, R. and Morrison, J. (1990), "Asymptotic analysis of random walks on a hypercube with many dimensions," *Random structures and algorithms*, 1, 52–72.

[8] Diebolt, J. and Robert, C.P. (1994), "Estimation of Finite Mixture Distributions by Bayesian Sampling," *Journal of the Royal Statistical Society*, B, 56, 363–375.

[9] Feller, W. (1968), *An Introduction to Probability Theory and Its Applications – Vol. II*, Wiley.

[10] Gelfand, A.E., and Smith, A.F.M. (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 86, 398–409.

[11] Gelman, A. and Rubin, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, no. 4, 457–511.

[12] Guihenneuc-Jouyaux, C. and Robert, C.P. (1998), "Finite Markov chain convergence results and MCMC convergence assessment", *Journal of the American Statistical Association*, to appear.

[13] Kemeny, J.G. and Snell, J.L. (1960), *Finite Markov Chains*, Springer-Verlag, New York.

[14] Lezaud, P. (1998), "Chernoff bound for finite Markov chains", *Ann. Applied Proba.* (to appear).

[15] Mann, B. (1996), *Berry-Esseen Central Limit Theorem for Markov Chains* PhD dissertation, Harvard University.

[16] Raftery, A.E., and Lewis, S. (1992), "How many iterations in the Gibbs Sampler?," in *Bayesian Statistics*, J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith (eds.), 4, 763–773, Oxford University Press, Oxford.

[17] Raftery, A.E., and Lewis, S. (1996), "Implementing MCMC," in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S.T. Richardson and D.J. Spiegelhalter (eds.), pp. 115–130, Chapman and Hall, London.

[18] Robert, C.P. (1996), *Méthodes de Monte Carlo par Chaînes de Markov*, Economica, Paris.

[19] Shapiro, S.S., and Wilk, M.B. (1965) "An analysis of variance test for normality", *Biometrika* **52**, 591–611.

[20] Seoh, M. and Hallin, M.(1997), "When does Edgeworth beat Berry and Esséen?," *Journal of Statistical Planning and Inference*, (to appear).

[21] Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions (with discussion)," *Annals of Statistics*, 22, 1701–1762.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399